

SCIENTIFIC REPORTS



OPEN

Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*

Seydina Issa Diop¹, Andrew D. W. Geering², Françoise Alfama-Depauw¹, Mikaël Loaec¹, Pierre-Yves Teycheney³ & Florian Maumus¹

Endogenous viral elements (EVEs) are viral sequences that are integrated in the nuclear genomes of their hosts and are signatures of viral infections that may have occurred millions of years ago. The study of EVEs, coined paleovirology, provides important insights into virus evolution. The *Caulimoviridae* is the most common group of EVEs in plants, although their presence has often been overlooked in plant genome studies. We have refined methods for the identification of caulimovirid EVEs and interrogated the genomes of a broad diversity of plant taxa, from algae to advanced flowering plants. Evidence is provided that almost every vascular plant (tracheophyte), including the most primitive taxa (clubmosses, ferns and gymnosperms) contains caulimovirid EVEs, many of which represent previously unrecognized evolutionary branches. In angiosperms, EVEs from at least one and as many as five different caulimovirid genera were frequently detected, and florendoviruses were the most widely distributed, followed by petuviruses. From the analysis of the distribution of different caulimovirid genera within different plant species, we propose a working evolutionary scenario in which this family of viruses emerged at latest during Devonian era (approx. 320 million years ago) followed by vertical transmission and by several cross-division host swaps.

Although the field of viral metagenomics is rapidly expanding the repertoire of viral genome sequences available for evolutionary studies¹, it only provides a picture of viral diversity over a very short geological time scale. However, viruses can leave molecular records in the genomes of their hosts in the form of endogenous viral elements (EVEs). EVEs are viral sequences that have been inserted in the nuclear genomes of their hosts by either active or passive integration mechanisms and in many cases have been retained for extended periods of time, sometimes millions of years. Although many EVEs are subject to sequence decay, it is common to be able to reconstruct ancestral viral genome sequences, particularly for high copy number EVEs, due to the random nature of the mutations that do occur. The study of EVEs does allow the evolution of viruses to be traced, much like a fossil record². For example, the study of endogenous retroviruses has uncovered the extended diversity and host range of retroviruses, and has provided evidence that they have a marine origin, and that they developed in parallel with their vertebrate hosts more than 450 million years ago (MYA)³⁻⁵.

Plant EVEs were first discovered a little more than 20 years ago⁶ but have only received a fraction of the research attention directed towards EVEs in humans and other animals. Most characterized plant EVEs are derivatives of viruses in the family *Caulimoviridae*⁷. The *Caulimoviridae* is one of the five families of reverse-transcribing viruses or virus-like retrotransposons that occur in eukaryotes⁸, and is the only family of viruses with a double-stranded DNA genome that infects plants (<https://talk.ictvonline.org/>). Unlike retroviruses, members of the *Caulimoviridae* do not integrate in the genome of their host as part of their replication cycle. However, caulimovirid DNA can be captured in the host genome, probably by illegitimate recombination mechanisms. In fact, five of the eight officially recognized genera of the *Caulimoviridae* have endogenous counterparts in at least one plant genome^{7,9}.

Recently, Geering *et al.*¹⁰ showed that EVEs from an additional tentative genus of the *Caulimoviridae*, called 'Florendovirus', are widespread in the genomes of cultivated and wild angiosperms, and provided evidence for the oldest EVE integration event yet reported in plants, at 1.8 MYA¹⁰. Sister taxa relationships between

¹URGI, INRA, Université Paris-Saclay, 78026, Versailles, France. ²Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, GPO Box 267, Brisbane, Queensland, 4001, Australia. ³UMR AGAP, CIRAD, INRA, SupAgro, 97130, Capesterre Belle-Eau, France. Correspondence and requests for materials should be addressed to F.M. (email: florian.maumus@inra.fr)

florendoviruses in South American and Australian plants suggested Gondwanan links and a minimum age of 34 MYA for this virus genus based on estimates of when land bridges between these two continents were severed. About 65% of all angiosperm species that were examined contained endogenous florendoviruses and for five, these sequences constituted more than 0.5% of the total plant genome content. Another remarkable feature of the biology of florendoviruses was their extremely broad host range, which included both the primitive ANITA-grade angiosperms, as well as the more advanced mesangiosperms. Putative bipartite florendoviral genomes were also found, a unique characteristic for any viral retroelement.

Almost contemporaneously to the study of Geering *et al.*¹⁰, Mushegian and Elena⁹ interrogated plant genomes for the presence of 30 K viral movement protein (MP) homologues, a superfamily of proteins that is common to the *Caulimoviridae* and a diverse range of plant viruses with single-stranded RNA genomes. While some of the MPs that were detected were undoubtedly those of endogenous florendoviruses and other recognized taxa, some MPs appeared to derive from previously undescribed genera in the *Caulimoviridae*. In this and a subsequent study¹¹, caulimovirid MPs were shown to be nearly universally distributed in angiosperm genomes but also, importantly, in the genomes or transcriptomes of conifers, ferns and club mosses, thus extending the host range of this family of viruses to non-flowering land plants.

The reverse transcriptase (RT) domain is the most conserved domain in the genome of viral retroelements and is used for classification^{12,13}. The strong sequence conservation of this domain allows high quality alignments to be generated, even for distantly related taxa, and therefore searches for homologues in plant genomes is the most sensitive and informative method to assess whether endogenous caulimovirids occur in the first place, and then what diversity exists. In the following study, we provide evidence that almost every vascular land plant (Tracheophyta), even the most primitive species such as the club mosses (Lycopodiophyte), contain endogenous caulimovirids. By analyzing the distribution of different genera, we unveil a complex pattern of associations and propose a scenario in which the *Caulimoviridae* would have emerged approximately 320 million years ago.

Results

Augmenting the diversity of known endogenous caulimovirids. A collection of RT domain sequences from known exogenous and endogenous caulimovirids was used to search for related sequences across the breadth of the Viridiplantae (four green algae, one moss, one club moss, four gymnosperms, and 62 angiosperms; Supplementary Table 1) using tBLASTn. Initially, over 8,400 protein-coding sequences were retrieved, all containing an RT domain with a best reciprocal hit against members of the *Caulimoviridae*, as opposed to the closely related *Metaviridae* (Ty3/Gypsy LTR retrotransposons). To provide a preliminary classification, sequences with at least 55% amino acid identity to each other were clustered and then iteratively added to our reference set of RT domain sequences to build a sequence similarity network. The successive networks were examined manually and representative sequences from each cluster were kept only when creating substantially divergent branches so as to cover an extended diversity of caulimovirid RTs with a core sequence assortment. While this network-based approach cannot be taken as phylogenetic reconstruction, it provided a practical method to explore diversity.

In the final sequence similarity network (Fig. 1), 17 groups with deep connections were identified, hereafter referred to as operational taxonomic units (OTUs). Remarkably, nine of these OTUs were distinct from recognized genera of the *Caulimoviridae*. Four of these novel OTUs were exclusively composed of sequences from gymnosperms, thereby confirming a significant host range extension for the *Caulimoviridae*. These OTUs were named Gymnendovirus 1 to 4. Two other novel OTUs were composed of RTs from various angiosperms and were named Xendovirus and Yendovirus. The last three novel OTUs were less populated, comprising sequences from one or two plant species (*Petunia inflata* and *Petunia axillaris*; *Vitis vinifera*; *Glycine max*; named species-wise: Petunia-, Vitis-, and Glycine-endovirus). This initial search therefore uncovered a significantly augmented diversity of caulimovirid RTs.

Endogenous caulimovirid RT (ECRT) density across the Viridiplantae. To perform a more comprehensive search for ECRTs in our collection of plant genomes, we used the sequences from the final sequence similarity network (Fig. 1) to search for ECRT nucleotide sequences that do not necessarily retain uninterrupted open reading frames. Using tBLASTn, we detected 14,895 genomic loci representing high-confidence ECRT candidates. Remarkably, ECRTs were found in nearly all seed plants, ranging from gymnosperms (ginkgo and conifers) to angiosperms. Over one-thousand ECRTs were detected in each of the genome assemblies of the gymnosperms *Picea glauca* (white spruce) and *Pinus taeda* (loblolly pine), as well as from the solanaceous plant species *Capsicum annuum* (bell pepper) (Fig. 2A). In general, we observed a positive correlation between plant genome size and the number of ECRTs, although there were notable exceptions, such as the monocot *Zea mays* (maize), which has a relatively large genome at 2.1 Gb but no detectable ECRT. Five other seed plants from our sample also lacked ECRTs, including two other monocots (*Zostera marina* and *Oryza brachyantha*) and three dicots in the order Brassicales (*Arabidopsis thaliana*, *Schrenkiella parvula* and *Carica papaya*). When the number of ECRTs was normalized against genome size, *Citrus sinensis* (sweet orange) and *Ricinus communis* (castor bean) had the highest densities at 2.3 and 2 ECRTs per Mb, respectively (Fig. 2B). The primitive ANITA-grade angiosperm *Amborella trichopoda* also had a relatively high density of ECRTs (1 ECRT per Mb) compared to an average density of 0.2 ECRT per Mb across the 62 seed plant species that were examined.

Caulimovirid RTs also detected in ferns and a clubmoss. From the plant genomes examined thus far, ECRTs were detected in gymnosperm genomes but not in those from the spikemoss *Selaginella moellendorffii* and from the moss *Physcomitrella patens*, which belong to the more basal land plant divisions Lycophyta and Bryophyta, respectively. Monilophytes, including ferns (class Polypodiopsida), are a sister lineage to seed plants^{14,15}, but no high quality genome assemblies are publicly available for these plants. However, six fern genomes have recently been sequenced at low coverage (approximately 0.4 to 2x genome size equivalent¹⁶ and we

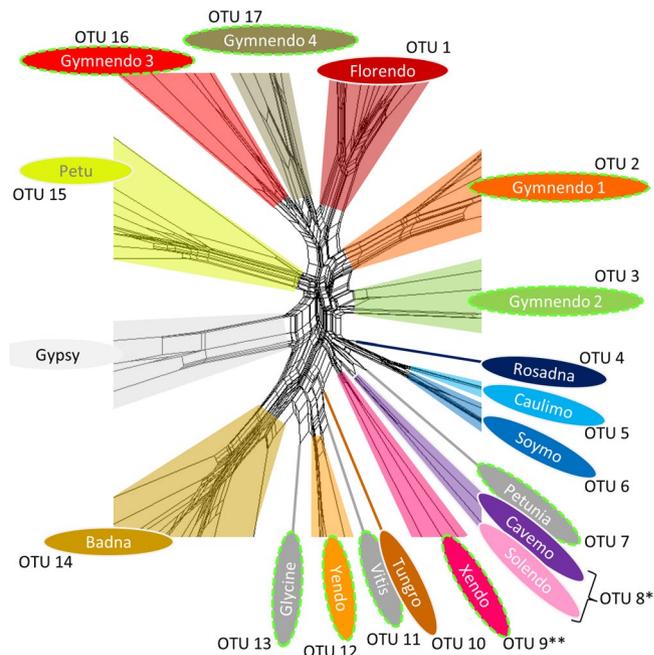


Figure 1. Augmented diversity of the *Caulimoviridae*. Core of a sequence similarity network constructed using an alignment of amino acid reverse transcriptase (RT) sequences from reference genera, representative endogenous caulimovirids and Ty3/Gypsy LTR retrotransposons. The full network is available in Supplementary Fig. 1. Operational taxonomic units (OTUs) that do not correspond to recognized genera are highlighted by dashed lime green outlines (referred to as novel OTUs). Each fill color corresponds to a different OTU, except for novel OTUs that contain only one type of sequence, which are shaded dark grey and named after the host plant genome of origin (Petunia-, Vitis-, and Glycine-virus). *RT clustering at 55% amino acid identity groups, which has led to the lumping of the genera *Cavemovirus* and *Solendovirus* in a single OTU (OTU 8). **Sequences grouped in the Xendovirus OTU are paraphyletic after phylogenetic reconstruction (see Fig. 3).

therefore screened these datasets for the presence of ECRTs. A total of twenty-one protein-coding ECRTs were detected in genomic contigs from five of the six fern species examined (Supplementary Table 1). Sequence similarity network reconstruction using representative fern ECRTs revealed that they form two novel OTUs that were named Fernendovirus 1 and 2 and numbered OTU #18 and 19, respectively (Supplementary Fig. 2).

Additional basal lineages of the Viridiplantae are represented in the 1,000 plant transcriptomes generated by the 1KP initiative^{17,15}. From this dataset, we found two transcript contigs (2.4 and 2.8 kilobases long, respectively) in the fern *Botrypus virginianus* (identifier BEGM-2004510) and *Lindsaea linearis* (identifier NOKI-2097008), which contained ECRTs (Supplementary File 1). Remarkably, we identified one more transcript contig (identified as ENQF-2084799, 2 kb) that contained an ECRT in the clubmoss *Lycopodium annotinum*, which belongs to the *Lycopoda*, the most basal radiation of vascular plants (Tracheophyta). It is not possible to determine whether the mRNAs were transcribed from exogenous viruses or from EVEs.

Phylogenetic reconstruction. Complete or near complete viral genomes were reconstructed from each novel OTU except Fernendovirus 1 and 2 (Supplementary file 1). From the fern genomic data sets, we were able to reconstruct fragments of the Fernendovirus 1 and 2 genomes of sufficient size for phylogenetic analysis. We also used the complete genomes of the type species of the eight currently recognized genera in the family *Caulimoviridae* (*Badnavirus*, *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Solendovirus*, *Soymovirus* and *Tungrovirus*), those of two unassigned viruses, Blueberry fruit drop-associated virus (BFDaV¹⁸) and Rudbeckia flower distortion virus (RuFDV¹⁹), and EVEs from the tentative genera *Orendovirus*²⁰ and *Florendovirus*¹⁰. From this library of sequences, we aligned conserved protease, reverse transcriptase and ribonuclease H1 domains to build a maximum likelihood phylogenetic tree (Fig. 3). Importantly, all newly identified EVEs grouped within the *Caulimoviridae* with strong bootstrap support.

In agreement with previous studies¹⁰, the tree revealed two sister clades, hereafter referred to as clade A and B (Fig. 3). Clade A comprised sequences from representatives of Xendovirus and Yendovirus OTUs and from members of the genera *Caulimovirus*, *Soymovirus*, *Rosadnavirus*, *Solendovirus*, *Cavemovirus*, *Badnavirus*, *Tungrovirus* and *Orendovirus*, as well as RuFDV and BFDaV. The Xendovirus OTU was found to be polyphyletic, hence a new taxon, Zendovirus, was created to accommodate the EVE from *Fragaria vesca*, while the EVE from *Gossypium raimondii* (cotton) was retained in Xendovirus. The Yendovirus OTU, comprising the EVE from *Capsicum annuum* (bell pepper), fell in the subclade comprising bacilliform-shaped viruses in the genera *Badnavirus* and *Tungrovirus*. The reconstructed genomes from novel OTUs found in single dicot species (Petunia-, Vitis-, and Glycine-endovirus) were discarded from the phylogenetic reconstruction as they significantly weakened the robustness of the tree. However, they unambiguously fell within clade A (data not shown).

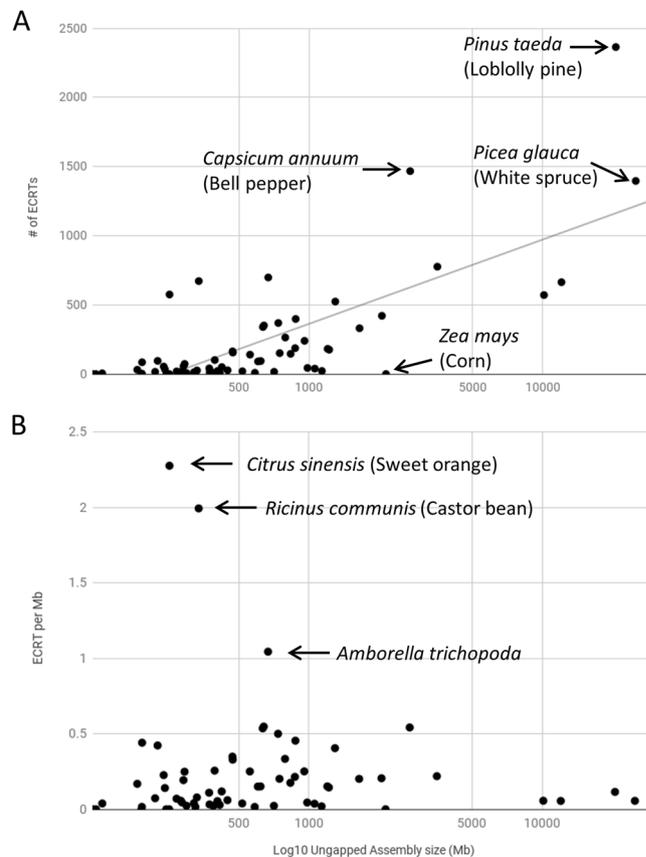


Figure 2. Estimates of copy numbers and densities of endogenous caulimovirids (ECRTs) in different plant genomes. **(A)** Number of ECRTs found in each plant genome as a function of Log_{10} genome size, expressed in megabases (assembly gaps excluded). The logarithmic trendline indicates moderate correlation between the number of ECRTs and genome size ($R^2 = 0.544$). **(B)** Density of ECRTs per megabase in each plant genome as a function of Log_{10} genome size, expressed in megabases (assembly gaps excluded). In **(A)** and **(B)**, arrows indicate outliers and corresponding plant species names.

Clade B comprised EVEs from the four gymnodovirus OTUs, the two fernendovirus OTUs, as well as representatives of the genus *Petuvirus* and the tentative genus *Florendovirus*. Gymnodovirus 2 EVEs were sister to all other clade B viruses, indicating that this group of viruses arose in gymnosperms. Interestingly, the angiosperm-infecting caulimovirids in this clade were polyphyletic, indicating independent origins and probable large host jumps of the most recent common ancestors. Fernendovirus 1 and 2 were monophyletic, and the club moss EVE placed within Fernendovirus 1. Again, the fernendoviruses appear to have arisen after a large host range swap of the most recent common ancestor.

ECRT distribution across seed plant genomes. To address the distribution of ECRTs in our collection of plant genomes, we determined the most likely position within the reference phylogenetic tree (Fig. 3) for the 14,895 sequences that we collected from using the pplacer program²¹. For this, we extracted sequences extending upstream and downstream of ECRT loci so as to retrieve contiguous fragments containing entire *pol* gene sequences (aspartic protease, RT and ribonuclease H1 domains). Using more relaxed length criteria, we extracted a total of 134 ECRT loci from the fern genomic data set that we also attempted to place on our reference tree.

Applying this strategy, we were able to classify a total of 13,834 ECRTs within specific OTUs (Fig. 4); the remaining ECRTs were placed on inner nodes of the reference tree. Overall, we observed striking differences between caulimovirid genera for both the number of ECRT loci and the number of plant species in which they were found. For instance, florendoviral RT loci were the most abundant, amounting to an overall total of 5,000 copies, and they were also found in the largest number of host species (46 of the 62 seed plant species that were screened). Petuviral RT loci were also well represented, with an overall total of 1,900 copies found in a total of 27/62 seed plant species, especially dicots. Among the novel OTUs, RTs classified as Yendovirus were found in the largest number of species, including monocots and dicots (Fig. 4).

Most importantly, the analysis of distribution of ECRTs in plant genomes revealed striking differences between ferns, gymnosperms and angiosperms (Fig. 4). No single OTU spans more than one plant division. Fernendovirus 1 sequences were exclusively distributed in fern and club moss genomes or transcriptomes. Similarly, ECRT loci that were assigned to one of the four Gymnodovirus OTUs were confined to gymnosperms. The three conifer genomes contained a mixture of ECRTs from the four Gymnodovirus OTUs but only ECRTs that were classified as Gymnodovirus 2 were detected in *Ginkgo biloba* (*Ginkgoales*). Within angiosperms, there was a dichotomy in

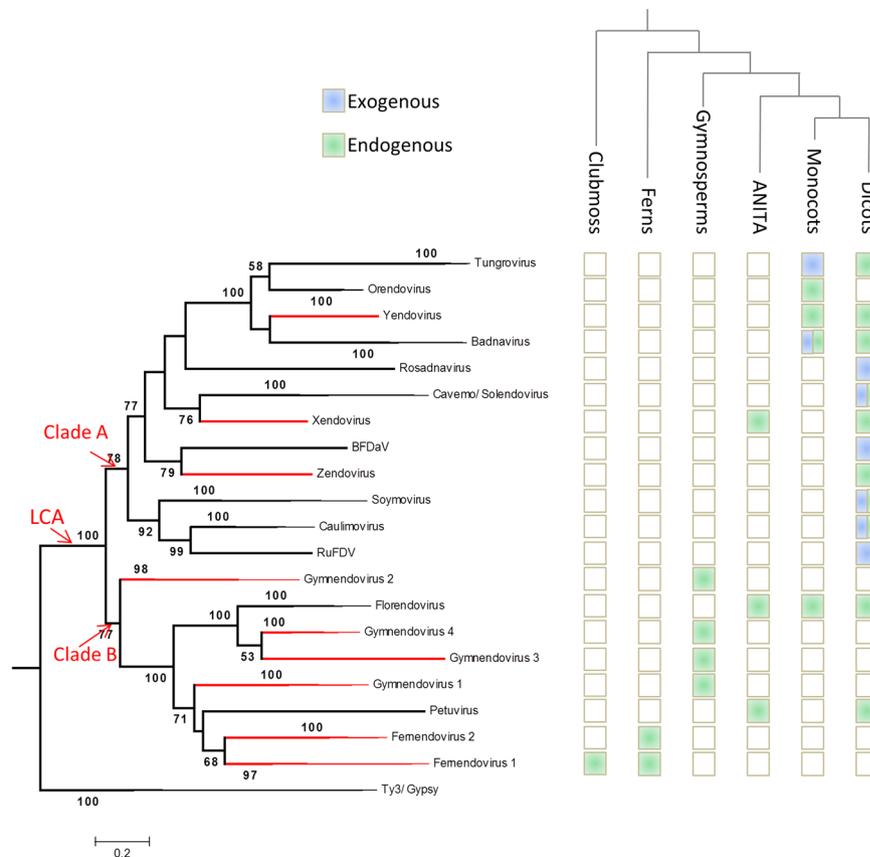


Figure 3. Phylogeny of the *Caulimoviridae*, as inferred using maximum likelihood criteria and a multiple sequence alignment of protease, reverse transcriptase and ribonuclease H1 domain sequences from recognized (black) and putative (red) genera. Ty3/Gypsy LTR retrotransposons were designated as the outgroup. Bootstrap support values below 50% are not shown. Branch nodes were collapsed until only taxa at the genus level were illustrated. Viruses included in the analysis were as follows: *Orendovirus* - *Aegilops tauschii* virus and *Brachypodium distachyon* virus; *Tungrovirus* - *Rice tungro bacilliform virus* (Type and West Bengal isolates); *Badnavirus* - *Commelina yellow mottle virus* and *Banana streak OL virus*; *Yendovirus* - *Capiscum annum* virus; *Zendovirus* - *Fragaria vesca* virus; Unassigned - Blueberry fruit drop-associated virus (BFDaV); *Caulimovirus* - *Cauliflower mosaic virus* and *Figwort mosaic virus*; Unassigned - *Rudbeckia flower distortion virus* (RuFDV); *Soymovirus* - *Soybean chlorotic mottle virus* and *Peanut chlorotic streak virus*; *Solendovirus* - *Sweet potato vein clearing virus* and *Tobacco vein clearing virus*; *Cavemovirus* - *Cassava vein mosaic virus* and *Sweet potato collusive virus*; *Petuvirus* - *Petunia vein clearing virus*; *Rosadnavirus* - *Rose yellow vein virus*; *Florendovirus* - *Fragaria vesca* virus and *Mimulus guttatus* virus; *Gymmendovirus 1* - *Pinus taeda gymmendovirus 1* and *Picea glauca gymmendovirus 1*; *Gymmendovirus 2* - *Pinus taeda gymmendovirus 2*, *Picea glauca gymmendovirus 2* and *Ginkgo biloba gymmendovirus 2*; *Gymmendovirus 3* - *Pinus taeda gymmendovirus 3*; *Gymmendovirus 4* - *Pinus taeda gymmendovirus 4* and *Picea glauca gymmendovirus 4*; *Fernendovirus 1*: *Cystopteris protrusa fernendovirus 1 contig 1*, and the transcript scaffolds BEGM-2004510 from *Botrypus virginianus*, NOKI-2097008 from *Lindsaea linearis*, and ENQF-2084799 from *Lycopodium annotinum*; *Fernendovirus 2* - *Dipteris conjugata fernendovirus 2 Contigs 2, 4 and 1319*. Clade A and B and the last common ancestor (LCA) of the *Caulimoviridae* are indicated with red arrows. The upper cladogram indicates the evolutionary relationships between major classes of vascular plants. At the intersection between both trees, colored boxes indicate the presence of either endogenous (green) or exogenous (blue) representatives of the *Caulimoviridae*.

the distribution of OTUs between monocots and dicots. On the one hand, *Yendovirus*, *Badnavirus*, *Orendovirus* and *Florendovirus* ECRTs are common in monocots, with the tentative genus *Orendovirus* being the only one specific to monocots. On the other hand, *Petuvirus*, *Florendovirus*, *Xendovirus*, *Cavemovirus/Solendovirus* and *Yendovirus* ECRTs are widely distributed in dicots and *Florendovirus* and *Yendovirus* ECRTs are additionally found in monocots.

Discussion

Endogenous viral elements are considered relics of past infections, and an extrapolation of the results from this study is that nearly every tracheophyte plant species in the world has at some point in its evolutionary history been subject to infection by at least one member of the family *Caulimoviridae*, as previously proposed by Mushegian and Elena⁹. This finding attests to the tremendous adaptability of the *Caulimoviridae* and also the large influence they have had on plant evolution, either as pathogens or as donors of novel genetic material to the plant genome.

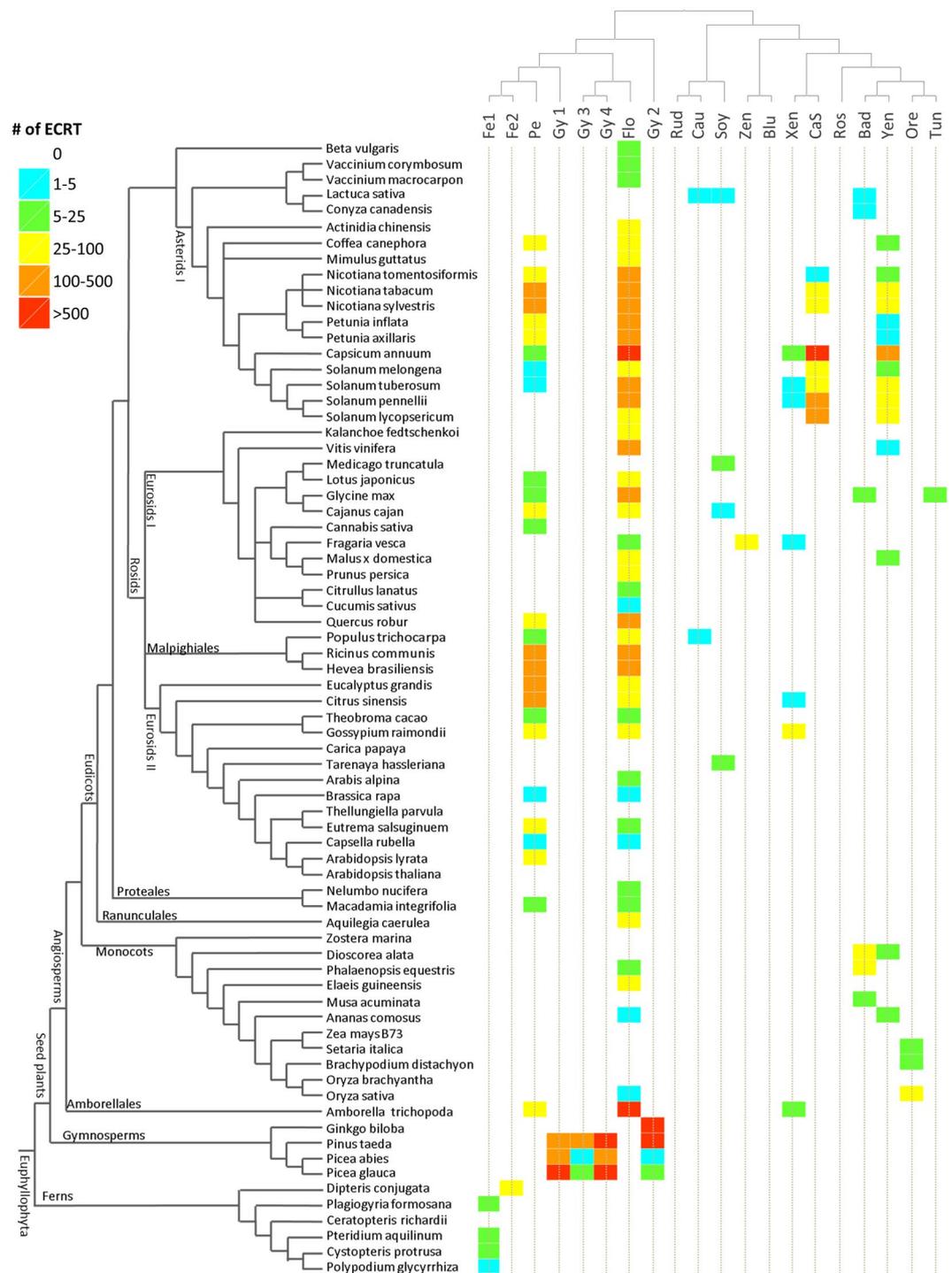


Figure 4. Distribution of endogenous caulimovirid RTs (ECRTs) among the Euphylllophytes. The cladogram on the left margin represents the phylogeny of euphylllophyte species investigated in this study; the names of major branches and nodes are indicated. The cladogram on the upper margin, which represents virus phylogeny, is derived from Fig. 3. At the intersection of these two trees, the colored boxes indicate the number of ECRT loci from each virus genus in each plant genome. Abbreviations of virus genera are as follows: Pe (*Petuvirus*), Gy1 (*Gymnendovirus* 1), Gy2 (*Gymnendovirus* 2), Gy3 (*Gymnendovirus* 3), Gy4 (*Gymnendovirus* 4), Fe1 (*Fernendovirus* 1), Fe2 (*Fernendovirus* 2), Flo (*Florendovirus*), Soy (*Soymovirus*), Rud (*Rudbeckia* flower distortion virus), Cau (*Caulimovirus*), Blu (Blueberry fruit drop-associated virus), Zen (*Zendovirus*), Xen (*Xendovirus*), Yen (*Yendovirus*), CaS (*Cavemovirus* + *Solendovirus*), Ros (*Rosadnavirus*), Bad (*Badnavirus*), Tun (*Tungrovirus*), Ore (*Orendovirus*).

The development of multicellularity in plants was likely a critical event in the emergence of the *Caulimoviridae*. Plants with rigid, cellulosic cell walls depend on plasmodesmata for the exchange of macromolecules²². The presence of a 30 K MP is an important feature of the *Caulimoviridae* that distinguishes it from the LTR retrotransposon family *Metaviridae*. This protein is crucial for the formation of systemic infection by allowing intercellular trafficking and phloem-loading of the viruses through increasing the size exclusion limit of plasmodesmata^{22,23}. The most recent common ancestor of the *Caulimoviridae* possibly arose after a recombination between a LTR-retrotransposon and an ssRNA virus with a 30 K MP, and this new hybrid virus would have been able to colonize the Tracheophyta but not necessarily more primitive plant lineages due to differences in plasmodesmatal biology. Although algae contain plasmodesmata, which superficially resemble those of higher plants, they are not homologous structures²⁴. Multicellular algae in the Charales, the closest relatives of land plants to have plasmodesmata, lack desmotubules that are continuous with the endoplasmic reticulum (ER), an anatomical feature that would likely limit intercellular trafficking of viruses that utilize the 30 K MP pathway. Furthermore, the true mosses (Bryophyta) do not have orthologues of the cysteine-rich receptor-like plasmodesmata-localized proteins that modulate cell trafficking in eudicots²².

The caulimovirid tree contained two major clades: clade A, comprising viruses that were found exclusively in mesangiosperm species and clade B, comprising viruses associated with all the major classes of Tracheophyta. Assuming the monophyly of clades A and B, the obtained plant-virus associations could be explained by the emergence of these viruses in a common ancestor of the gymnosperms and angiosperms followed by several major host swaps: in clade A, between monocots and dicots, and in clade B between gymnosperms and angiosperms in the case of florendoviruses and petuviruses (although the position of this latter genus in the tree is uncertain), and from gymnosperms to ferns and clubmoss in the case of fernendoviruses. Following this scenario, the *Caulimoviridae* could have emerged at the latest with the Spermatophyta, *i.e.* during the Devonian era, about 320 MYA²⁵.

When recapitulating the distribution of EVEs in plant genomes, the known host range of exogenous viruses, and the phylogenetic relationships between caulimovirid OTUs and major groups of vascular plants (Fig. 3) to infer the evolutionary trajectories of plant-virus coevolution, we obtain a complex pattern of host-virus associations. At the OTU level, the host distributions of petu- and xendovirus, including dicots and the ANITA grade angiosperm (*Amborella trichopoda*) but not any of the monocot species, is suggestive of horizontal transfer. In addition, although vertical transmission is overall well supported by a co-evolutionary study of florendoviral EVEs and their host species¹⁰, it could not be confirmed for *A. trichopoda*. Together with the observation that *A. trichopoda* presents a high density of ECRTs (Fig. 2B), this may suggest that this species is permissive to infection by a range of caulimovirid genera and/or that it represents a hotbed for the emergence of caulimovirid genera, some of which swapped towards mesangiosperms. We however cannot rule out the possibility that petu- and xendovirus were lost in monocots.

Methods

Discovery and clustering of novel caulimovirid OTUs. We built a library containing an assortment of amino acid (aa) sequences from 54 RT domains including four from *Retroviridae*, six from Ty3/Gypsy LTR retrotransposons, 41 from eight different genera of *Caulimoviridae* (Florendovirus, *Caulimovirus*, *Tungrovirus*, *Cavemovirus*, *Solendovirus*, *Badnavirus*, *Soymovirus*, and *Petuvirus*), two from *Picea glauca*, and the one from the DIRS-1 element. We compared this library to a collection of 72 genome assemblies from the Viridiplantae (listed in Supplementary Table 1) using tBLASTn with default parameters (except $-e$ option set to $1e^{-5}$). The hit genomic loci were merged when overlapping and their coordinates were extended 120 bases upstream and downstream. Extended hit loci were translated and the protein sequences of length ≥ 200 aa were compared to the initial RT library using BLASTp with default parameters (except $-e = 1e^{-5}$). Queries with best alignment score against *Caulimoviridae* over at least 170 residues were selected for further analysis. For each plant species, the selected set of RT aa sequences were clustered following sequence similarity using the UCLUST program²⁶ with identity threshold set at 80%. The longest sequence from each resulting cluster was considered as the representative sequence and it was appended to the initial RT library. To detect potential false positives, each set of sequences (each consisting of the initial RT library and cluster representatives from one species) was aligned using MUSCLE followed by filtering of lower fit sequences using two rounds of trimAl v1.2²⁷ to remove poorly aligned sequences ($-\text{resoverlap } 0.75\text{--seqoverlap } 50$) separated by one round to remove gaps from the alignment ($-\text{gt } 0.5$). The representative sequences from each plant species that passed this selection were combined into a single file and appended to the initial RT library to be clustered with UCLUST using an identity threshold of 55%. At this level of similarity, aa RT sequences from every genus in the *Caulimoviridae* was separated except those from *Cavemovirus* and *Solendovirus*.

Starting with the first cluster, one or more sequences presenting high quality alignment and containing several conserved residues as determined contextually for each cluster were then manually selected to be representative of the diversity observed within each cluster. The following clusters were processed similarly while keeping the representative sequences selected from previously processed clusters. Clusters containing ECRT sequences from only one plant species were analyzed only when they contained at least three sequences. After processing each cluster individually, a total of 56 ECRT sequences detected here and 20 RT from known genera were selected for their remarkable divergence. Together with four RT sequences from Ty3/Gypsy LTR retrotransposons, these combined sequences (hereafter referred to as “diverse library”) were aligned with the GUIDANCE2²⁸ program using MAFFT²⁹ to generate bootstrap supported MSA and to remove columns ($-\text{colCutoff}$) with confidence score below 0.95 (16/244 columns removed in the RT sequence from CaMV). The resulting MSA was then used to build the sequence similarity network shown in Fig. 1 and Supplementary Fig. 1 with SplitsTree4³⁰ applying the NeighborNet method with uncorrected P distance model and 1,000 bootstrap tests. Manual analysis of this network enabled the discrimination of 17 distinct groups sharing deep connections among caulimovirid sequences.

In response to the discovery of several novel OTUs, we repeated ECRT mining in plant genomes using the diverse library as query. This second search was also designed to be more sensitive as it took into account DNA sequences instead of uninterrupted ORFs. The workflow was identical to the one employed for the initial search until obtaining the set of extended hit loci. These were directly compared to the diverse library using BLASTx with default parameters (except $-e = 1e-5$). Queries with a best alignment score against any member of the *Caulimoviridae* with an alignment length above 80% of subject length (set generically to 576 bp considering an average size of RT domains of 240 aa) were selected for phylogenetic placement.

Phylogenetic analysis. Fragments of virus sequence were assembled using CodonCode aligner 6.0.2 using default settings or using VECTOR NTI Advance 10.3.1 (Invitrogen) operated using default settings, except that the values for maximum clearance for error rate and maximum gap length were increased to 500 and 200, respectively.

Phylogenetic reconstruction was performed using the contiguous nucleotide sequences corresponding to the protease, reverse transcriptase and ribonuclease H domains. Whole sequences of representatives of the different genera of the *Caulimoviridae* and Ty3/Gypsy LTR retrotransposons were first globally aligned using MAFFT v7.3²⁹. The core genomes were extracted and then locally re-aligned using MAFFT. The GTRGAMMA model of evolution was selected after examination of the alignment with pmodeltest v1.4 (from ETE 3 package³¹). Phylogenetic inference with maximum likelihood (ML) criteria was then performed using RaxML v8.2³² under the predicted model with 500 ML bootstrap replicates.

The resulting tree was then used as a reference to classify the ECRT loci mined from plant genomes. We first added query sequences from each plant species separately to the reference alignment and aligned each library using MAFFT v7.3 (with options `--addfragment`, `--keeplength` and `by reordering`). The most likely placement of each ECRT sequence on the reference tree was then tested using pplacer v1.1 alpha19²¹ with the option (`--keep-at-most 1`), which allows one placement to be kept for each query sequence. The python package Taxit was used to construct a reference package which we used to run pplacer.

Data availability. The datasets and scripts generated during the current study are available from the corresponding author on request.

References

- Roossinck, M. J. Deep sequencing for discovery and evolutionary analysis of plant viruses. *Virus Res.* **239**, 82–86 (2017).
- Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479–480**, 26–37 (2015).
- Hayward, A., Grabherr, M. & Jern, P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc. Natl. Acad. Sci. USA* **110**, 20146–20151 (2013).
- Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. USA* **112**, 464–469 (2015).
- Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun* **8**, 13954 (2017).
- Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc. Natl. Acad. Sci. USA* **93**, 759–764 (1996).
- Teycheney, P. Y. & Geering, A. D. In *Recent advances in plant virology*. (eds C. Caranta, M. A. Aranda, M. Tepfer & J. J. Lopez-Moya) 343–362 (Caister Academic Press, Norfolk; 2011).
- Pringle, C. R. The universal system of virus taxonomy of the International Committee on Virus Taxonomy (ICTV), including new proposals ratified since publication of the Sixth ICTV Report in 1995. *Arch. Virol.* **143**, 203–210 (1998).
- Mushegian, A. R. & Elena, S. F. Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology* **476**, 304–315 (2015).
- Geering, A. D. *et al.* Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5**, 5269 (2014).
- Mushegian, A., Shipunov, A. & Elena, S. F. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. *BMC Biol.* **14**, 68 (2016).
- Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362 (1990).
- Hansen, C. & Heslop-Harrison, J. S. Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. *Advances in Botanical Research* **41**, 165–193 (2004).
- Kenrick, P. The relationships of vascular plants. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 847–855 (2000).
- Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. USA* **111**, E4859–E4868 (2014).
- Wolf, P. G. *et al.* An Exploration into Fern Genome Space. *Genome Biol. Evol.* **7**, 2533–2544 (2015).
- Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
- Diaz-Lara A, Martin R. R. Blueberry fruit drop-associated virus: a new member of the family *Caulimoviridae* isolated from blueberry exhibiting fruit-drop symptoms. *Plant Disease* **100**, 2211–2214 (2016).
- Lockhart B, Molloy D, Olszewski N, Goldsmith N. Identification, transmission and genomic characterization of a new member of the family *Caulimoviridae* causing a flower distortion disease of *Rudbeckia hirta*. *Virus Research* **241**, 62–67 (2017).
- Geering, A. D., Scharaschkin, T. & Teycheney, P. Y. The classification and nomenclature of endogenous viruses of the family *Caulimoviridae*. *Arch. Virol.* **155**, 123–131 (2010).
- Matsen, F. A., Kodner, R. B. & Armbrust, E. V. Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
- Link, K. & Sonnwald, U. In *Plant-Virus Interactions: Molecular Biology, Intra- and Intercellular Transport*. (ed. T. Kleinow) 1–37 (Springer International Publishing, Cham; 2016).
- Rojas, M. R. *et al.* In *Current Research Topics in Plant Virology*. (eds A. Wang & X. Zhou) 113–152 (Springer International Publishing, Cham; 2016).
- Brunkard, J. O. & Zambryski, P. C. Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. *Curr. Opin. Plant Biol.* **35**, 76–83 (2017).
- Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).

26. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
27. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
28. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**, W7–14 (2015).
29. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
30. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
31. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
32. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

Acknowledgements

We thank Mark Tepfer for critical reading of a previous version of this manuscript. PYT was funded by the Guadeloupe Region and the European Regional Development Fund.

Author Contributions

S.D., A.D.W.G., and F.M. performed research; F.A-D. and M.L. provided the research environment; S.D., A.D.W.G., P.Y.T., and F.M. analyzed the data; A.D.W.G., P.Y.T., and F.M. wrote the manuscript with contributions from S.D.; F.M. designed the research.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-16399-x>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018